Practical Performance/Power Alternatives within an Existing CMOS Technology Generation

Kerry Bernstein, John E. Bertsch, William F. Clark, John J. Ellis-Monaghan, Larry G. Heller, Edward J. Nowak
IBM Microelectronics, 1000 River Road, Essex Junction, VT 05452
kbernstein@vnet.ibm.com

The tremendous demand for advanced microprocessors is making developers offer a superior product at a competitive price which meets specific performance, power, and reliability requirements. Today's products are designed and built with essentially the same design, fabrication, and test tools used by all industry manufacturers and are constrained by the same device physics, package impedance, heat dissipation capability, and battery energy density limitations. In addition, the current technology generation presents a new set of difficult challenges for the treatment of power dissipation.

Lowering voltage to reduce power diminishes FET overdrive and the performance of the device. Recapturing performance by migrating quickly to the next scaled technology generation, however, makes low-power designs expensive, and is not always necessary. Standard full scaling preserves physical and electrical relationships between parameters. On the other hand, selective scaling of specific device parameters may allow the exploitation of existing tools and designs by the use of a different design point on the process "surface." This paper describes recent attempts to explore the feasibility of selective scaling and anticipated constraints associated with future technology generations.

SELECTIVE SCALING

Conventional CMOS concepts of scaling vertical and horizontal device dimensions and the power supply voltage (V_{dd}), by a common factor are well documented [1]. With the exception of V_{dd} and threshold voltage (V_t), the principles of MOS scaling have historically been practiced by the industry through several technology generations. Efforts to obtain decreases in V_t with technology scaling, however, have been limited. Subthreshold (leakage) current in MOSFETs is due to weak inversion carriers, whose population density is proportional to the Boltzmann factor, e**-(phi)sub-s/kT, where k, T and (phi)sub-s are Boltzmann's constant, the temperature (absolute) and the silicon surface potential, respectively. Since (phi)sub-s is proportional to (V_g-V_t), decreasing V_t leads to exponentially increasing leakage current, thereby limiting the amount of V_t reduction possible. Long-lasting power supply voltage standards, such as the 5V standard, have discouraged the scaling of system-level power supplies. Additionally, high-performance circuit design requirements have limited the allowable reduction in device drive, V_{dd}-V_t, and will continue to do so [2].

As an alternative to full scaling, selective scaling exploits existing tool capabilities by finding new device operating points within the process tool window which are acceptable for the application. These alternate process settings allow operation at reduced voltages, thereby mitigating heat dissipation and battery life issues. Generally, specific device parameters may be identified which require the next generation of process tooling to be improved substantially, and so are ineligible to for selective scaling. These parameters include device length tolerance control, overlay/alignment control, image tolerance, and junction depths. Alternative design points can be explored by changing parameters that a current installation can achieve within its existing process window. These would include device threshold tailoring implants, source/drain junction simplification (i.e., elimination of grading), gate dielectric thickness, interconnect film thickness, and image photo/etch bias. The selection of parameters to scale and the magnitude of departure from convention are governed by the MAXIMUM tolerable noise, standby current, cost, power consumption, and by the MINIMUM tolerable reliability and performance. For products with an acceptable tradeoff on the "process surface," a potential exists for extended product life and market participation.

A recent experiment demonstrated that a 3.5X power reduction in a high-performance, 0.6μm CMOS product technology, could be achieved with only minor process changes and the same mask set [3]. For the IBM 3.6V PowerPC 601 microprocessor, an existing, well-characterized product was selectively scaled in thresholds and gate oxides. Gate oxide (t_{ox}) and NFET device thresholds were selected to achieve power/performance targets at reduced V_{dd} without changing masks and with only minor process changes. As a result, reliability exposures were minimized at that V_{dd} [4]. Elimination of the need for horizontal scaling allowed relatively quick and inexpensive implementation of the technique, compared to the retooling typically required by a full technology scaling.

To maintain functionality of the PowerPC 601 at reduced V_{dd}, V_t and t_{ox} were reduced at least as fast as V_{dd} for the NFET. Because load capacitance would not be scaled in this

ISLPED 1996 Monterey CA USA
0-7803-3571-8/96/$5.00 ©1996

experiment, we attempted to keep the MOSFET drain currents constant with (selective) scaling to retain performance with reduced V_{dd}. In full scaling, V_g, V_t, and V_{dsat} would scale in proportion to V_{dd} while oxide capacitance (C_{ox}) would scale inversely, keeping constant current per width. With selective scaling, V_{dsat} decreases more slowly than $V_{dd,}$ since L_{eff} is not reduced. Figure 1 compares full scaling to our selectively scaled approach. For a worst-case modelled analysis, the pinch-off voltage was assumed to remain constant with selective scaling cases. In reality there is some reduction. Performance, f, is expected to be

$$f/f_0 = (V_{dd0}/dd) \text{ x } (C_{load0}/C_{load}) \text{ x } I_{drive}/I_{drive0}) = 0.80 \text{ X.}$$

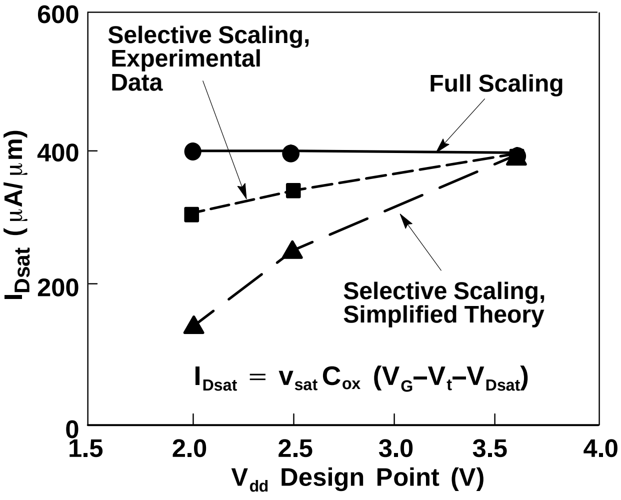

Figure 1. Drain current in an n-type MOSFET (with $V_{gs} = V_{ds} = V_{dd}$) vs. V_{dd} design point for full scaling compared to the selective scaling case shown.

Since PFET V_t was not reduced to preserve static switch point integrity, I_{drive} was reduced by approximately 20%. Active power reduction expected is:

$$P/P_0 = (V_{dd}/V_{dd0})^{**}2 \text{ x } (C_{load}/C_{load0}) \text{ x } (f/f_0) = 0.44x.$$

Power versus performance is shown in Figure 2 for traditional and selective scaling. Note that selective scaling improves power consumption while attempting to minimize the loss in performance. On the other hand, full scaling simultaneously improves both. The cost of full scaling, however, is new tooling.

Process changes were made to target 50% of V_{dd}. Because the PFET has a compensated channel, arsenic was substituted for phosphorus in the channel to maintain short-channel behavior in the scaled processes. In the 3.6V process, tox was reduced from 11.5nm to 4.9nm and 7.0nm for the 2.0V and 2.5V design points, respectively. Polysilicon gate-electrode depletion resulted in electrical equivalent gate-oxide thicknesses of 5.5nm and 7.5nm for the 2.0V and 2.5V cases, respectively. The NFET I_{dsat} achieved is also shown in Figure 1. Note that nearly constant I_{dsat} was achieved by reducing t_{eq} and V_t slightly faster than V_{dd}.

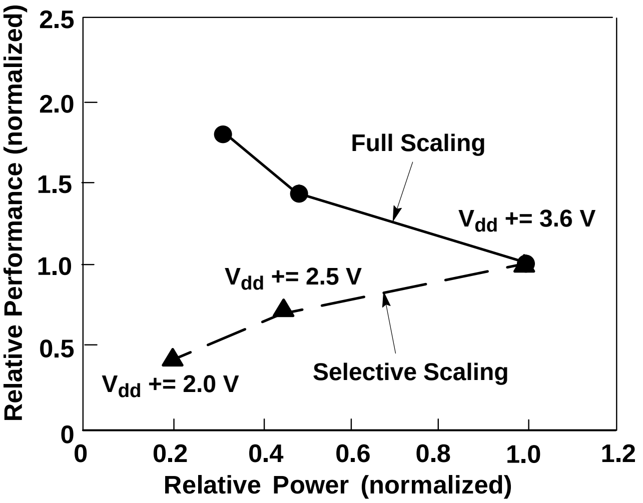

Figure 2. Performance (normalized to the 3.6V design case) vs. active power (also normalized to the 3.6V design case) for selective scaling (experimental data) vs. full scaling (theoretical expectation).

Table 1. IBM PowerPC 601 Characteristics.

Die size	10.95mm x 10.95mm
Performance	80MHz
Power consumption	8.5W
Device count	2.8 M
Signal I/O	184
Power supply	3.6 +/- 5% V

The circuit style on the vehicle, the IBM PowerPC 601 (Table 1), is predominantly static. This RISC processor has no dynamic domino or DCVS-style circuitry on board. Clock buffers and redrivers on the chip shape input clocks but do not run autonomously. There are no phase-locked loops on the product. A limited amount of ratioed logic circuitry is used. There were no alterations to the mask or the design for the experiment. Wafer level horizontal dimensions are identical to those of the standard product.

Performance of up to 68.4MHz was observed at room temperature for V_{dd} as low as 2.1V. The masks used were from a version of the product which achieved 80MHz performance at room temperature with standard processing. Standby currents, which on the standard product rarely exceed 100μA, were observed to be between 25 and 40mA in the experiment. Active power was found to be 2.0W on average while the standard production test patterns were running, as compared to 7.5W seen on standard production hardware at 80MHz. This is shown graphically in Figure 3. Functional module yield of experimental hardware was equivalent to that found on standard 3.6V production hardware. Except for input-drive and output-sense voltage levels, standard production criteria were used in testing the modules. Parts underwent standard 601 processing and received normal handling through wafer and module build, except in the well implant and gate-oxide growth sectors.

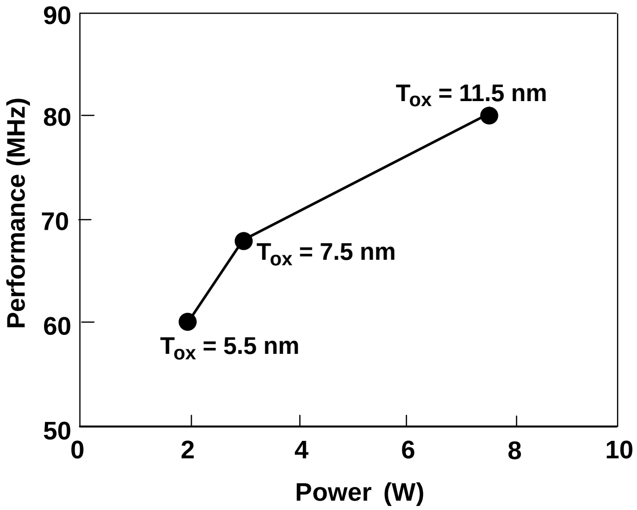

Figure 3. Performance vs. power for standard and experimental modules, as measured at their targeted operating voltages.

The reduced operating voltage enhanced the reliability associated with channel-hot-electron threshhold degradation by 2X and, as operating temperature was reduced, the interconnect electromigration-dependent lifetime was improved by 1.5X.

The PowerPC 601 chip is rich in NAND structures. Good static CMOS design technique exploits logic NANDs by stacking multiple NFETs in series to ground, rather than NORs which stack multiple PFETs to V_{dd}. In CMOS technology, NFETs have more than twice the current of PFETs. The decision not to modify PFET thresholds had only a minor impact on overall performance.

FUTURE TRADEOFFS

The magnitude of selectively scaled gate oxide and device threshold reduction possible in new technology generations is diminishing. Theoretically, the minimum threshold needed to perform CMOS logic is governed by thermodynamic considerations, and is near zero volts. In practice, the introduction of various coupling and noise sources requires V_t to be considerably higher. Dynamic logic, used increasingly for performance, presents additional sensitivities to low threshold. Modified thresholds must be selected to accommodate the loss of associated noise immunity. High subthreshold leakage is associated with poor charge retention in dynamic logic and can affect minimum allowable cycle time, burn-in functionality, and power-saving logic. Higher standby current also increases chip DC current, thereby reducing the interval for battery recharge.

The statistical line width variation within a chip limits further threshold voltage reduction by requiring added margin to offset resulting increased leakage currents. Given the expected variation at 0.18μm channel length, the leakage associated with a minimum threshold of 50-100mV under the planned minimum V_t must be anticipated.

REDUCED THRESHOLD EFFECTS

A popular precharged-design style, the dynamic domino, was selected to assess the effects of lowering threshold (Figure 4 inset) [5]. Dynamic dominos are known for their high performance, low input capacitance, and simplicity, but also for high noise generation and power consumption caused by clocking. Domino logic from a current 32-bit RISC microprocessor design was modelled using a Semiconductor Industry Association (SIA) 0.18μm technology projection, summarized in Table 2 [6].

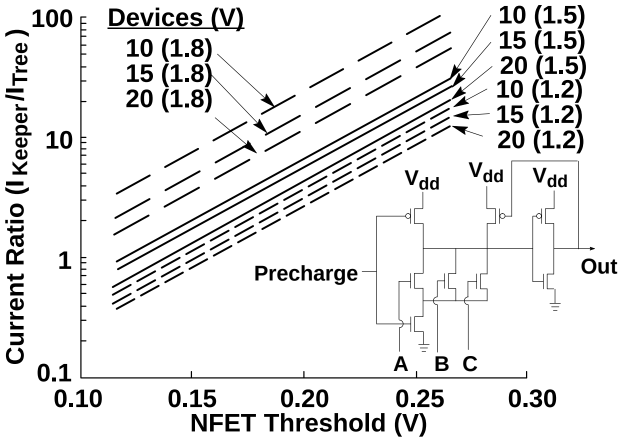

Figure 4. Sustainable leakage.

Table 2. SIA 0.18-μm technology.

T_{ox}	4.5nm	V_{tn}, V_{tp} (0.16μm)	0.21V
L_{eff}	0.14-0.18μm	M1 width/space	0.22/ 0.33μm
V_{dd}	1.2 - 1.8V	M1 thickness	0.55μm

Leakage in dynamic logic can unintentionally pull precharge nodes below their switchpoint during sampling. Lower thresholds also cause a reduction in the maximum allowable signal width of a logic book, with each device allowing higher subthreshold current to ground. Figure 4 shows the ratio of PFET replacement current to NFET leakage current for a wide-domino NOR, varying the number of pulldowns. For the extreme cases, the ratio remains above 1.00 for 15 or fewer pulldown devices, which indicates that, alone, leakage to ground is not a functionality issue for a moderately wide NOR employing reduced NFET thresholds at 100 degrees C for V_{dd} of 1.2 to 1.8V.

The composition of a current desktop microprocessor design was examined to evaluate chip DC standby current, or "quiescent Idd." Each component, i.e., SRAMs, registers, custom logic, book logic, I/O, was assessed for average PFET and NFET device size, likely logic state and the resulting leakage mode. The cumulative leaking NFET and PFET effective device widths in each category were then used to determine a total subthreshold current at 100 C for a range of thresholds and voltages. Only the thresholds of high-performance logic devices which can leverage increased overdrive were reduced (Figure 5).

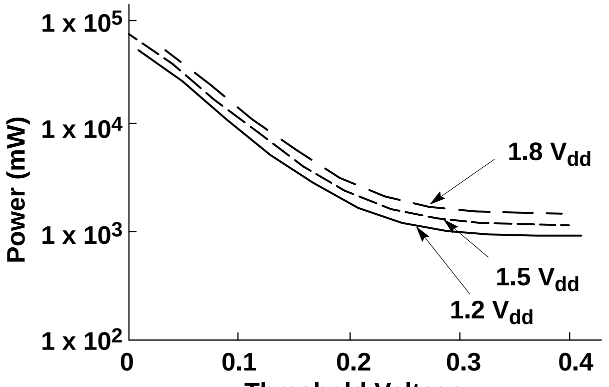

Figure 5. Chip standby power-threshold dependence.

Generally, DC noise margin is linearly related to device threshold. Figure 6 shows the results of modeling a logic path composed of reduced threshold dominos, measuring DC noise margin-low (NML) to the unity gain point for internal stages. For the design being considered, a noise immunity of 15% to 20% of V_{dd} is required to accommodate its ground bounce and capacitive interconnect coupling. Figure 6 indicates a minimum V_t of 145 to 218mV at 1.5V.

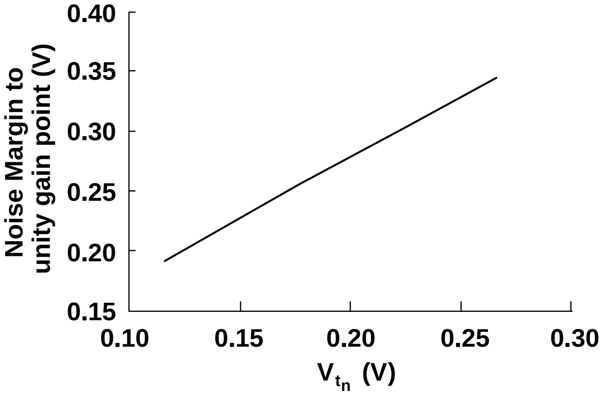

Figure 6. Noise-margin threshold-voltage dependence.

In addition to ground bounce and coupling, DI/dt noise, alpha particles, and charge sharing/division on multiple-level domino logic trees must be considered. It is assumed that these considerations are included in the design of the dynamic logic book. Of additional concern are systems which have inputs from components powered by higher supply voltages. The noise carried by those inputs will have a magnitude proportional to the source supply.

Allocating 75% of the noise budget to interconnect coupling noise, the allowable peak noise for a 1.5V V_{dd} supply is 169 to 225mV. The coupled noise for a quiescent line between two actively switching lines is shown in Figure 7. Approximately 350mV of noise margin is available at a threshold of 270mV, which translates to approximately 1 mm of allowable interconnect length. By reducing the threshold voltage to 170mV, the allowable noise margin degrades to 230mV which reduces the allowable interconnect length to 200μm.

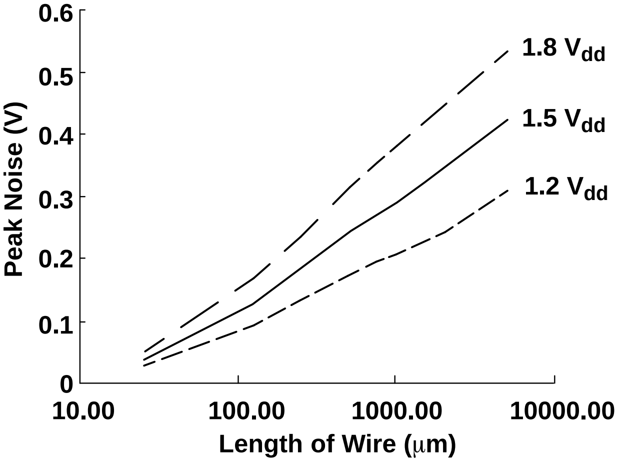

Figure 7. Interconnect-induced noise.

The motivation for reducing threshold is the recovery of overdrive lost to lower operating voltage. Figure 8 shows the impact on delay (per stage) of a dynamic path from the microprocessor being modelled, comprised of 125μm of interconnect between each of 10 stages. At 1.5V the minimum thresholds providing the required noise margin give a performance gain of up to 10%. Reducing threshold becomes more effective at increasing performance at lower operating voltages. The ability to control gate line-width tolerance in processing translates into additional performance by allowing lower nominal threshold selection.

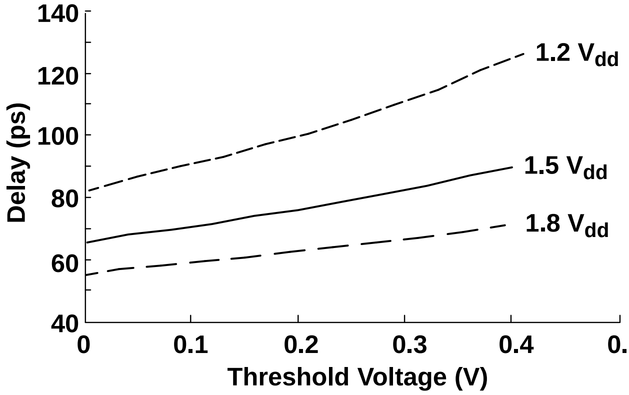

Figure 8. Performance dependence on threshold.

The energy-delay product shown in Figure 9, as a function of threshold, demonstrates a profound dependence on operating voltage and a weak dependence on threshold [7]. A 4X reduction in power-energy product was realized in the modelled processor, going from 1.8V V_{dd} to 1.2V V_{dd}. This leverage will be short-lived however. Figure 10 illustrates the difficulty in recapturing overdrive as V_{dd} becomes smaller.

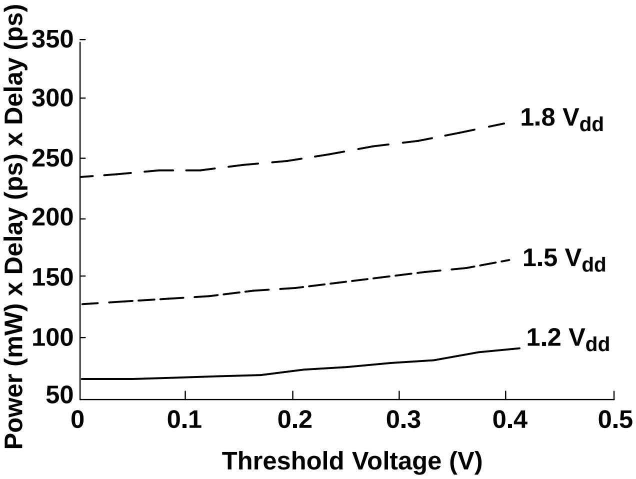

Figure 9. Energy-delay-product dependence on threshold.

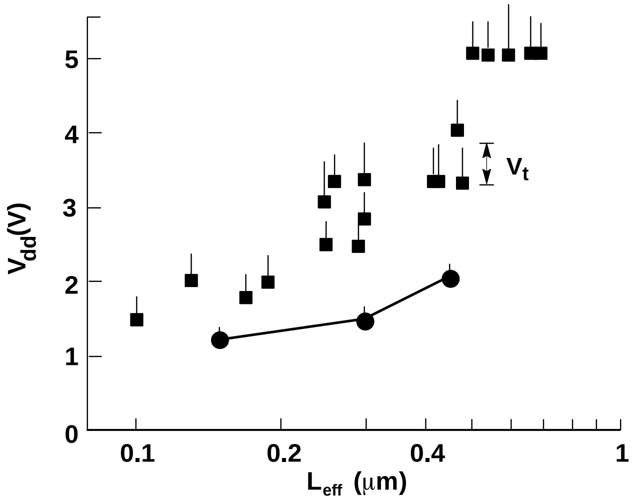

Figure 10. V_{dd} and V_t for recently published processes, and selective scaling. Convergence is apparent.

RESULTS

The data indicates that approximately 0.1 x V_{dd} is the practical lower bound on FET device threshold for a 0.18μm hypothetical process. At this setting, the designer realizes a performance improvement of up to 10% compared to a design implemented with conventionally scaled thresholds. Chip standby power is expected to increase by approximately 10% in the implementation described. In the absence of other special preparation, the global chip integrator must limit interconnect lengths on inputs of dynamic circuitry to approximately 400 μm. Circuit library elements with logic widths greater than 15 signals must also be avoided.

Over the long term, selective scaling of V_t and t_{ox} is confronted with the same limitation as conventional scaling; V_t reduction bottoms out at roughly 200mV in conventional CMOS logic schemes [2] due to standby power constraints. This, in turn, limits V_{dd} reduction when MOSFET performance is optimized for the active switching power constraint. As a result, we foresee power/performance-optimized CMOS converging on a V_{dd} floor in the neighborhood of 1.0V for power-constrained systems.

CONCLUSIONS

The use of selective scaling to extend the life of existing processes and products can be useful and has been demonstrated on a commercial product. However, the suitability of alternate operating points to the application must be examined. Both selectively scaled and fully scaled designs in the 0.18μm realm will need to consider interconnect length and logic bookwidth to maintain adequate noise immunity at lower thresholds. The window of opportunity for selective scaling is diminishing, however, for current CMOS design styles. The loss of overdrive sustained by partial scaling with low V_{dd} affects performance more than full scaling, caused by the lack of reduced capacitance. As lithography shrinks, selectively scaled and fully scaled/migrated designs are converging on the fundamental limitations that govern V_t reduction. This constraint may limit V_{dd} to no lower than 1.0V even for full scaling. New techniques will be required to supplant scaling to overcome this barrier.

ACKNOWLEDGEMENTS

The authors wish to thank R.Masleid, C.Durham, L.Su, J.Lary, S.Schmidt, and J.Hiltebeitel for constructive discussion and suggestions.

REFERENCES

1. R. H. Dennard, F. H. Gaensslen, H. Yu, V. L. Rideout, E.Bassous, A. R. Leblanc, “Design of ion-implanted MOSFET’s with very small dimensions,” IEEE J. Solid State Circuits SC-9, 256 (1974).

2. E. J. Nowak, “Ultimate CMOS ULSI performance,” IEDM Technical Digest, pp. 115-118 (1993).

3. J. E. Bertsch, K. Bernstein, L. G. Heller, E. J. Nowak, F. R. White, “Experimental 2.0V power/performance optimization of a 3.6V design CMOS microprocessor - PowerPC 601,” 1994 IEEE Symposium on VLSI Technology, 7A2 (1994).

4. D. Bouldin, “Reliability issues in multilevel interconnects,” 1994 Multilevel Interconnection Seminar Proceedings, University of South Florida, 5 (June,1994).

5. R. H. Krambeck, C. M. Lee, and H.F.S. Law, “High speed compact circuits with CMOS,” IEEE JSSC, Vol. SC-17, No. 3, June 1982, pp. 614-619.

6. Semiconductor Industry Association, “The national technology roadmap for semiconductors.”

7. A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, “Low-power CMOS digital design,” IEEE J. Solid-State Circuits 27, No. 4, pp. 473-484 (April 1992).